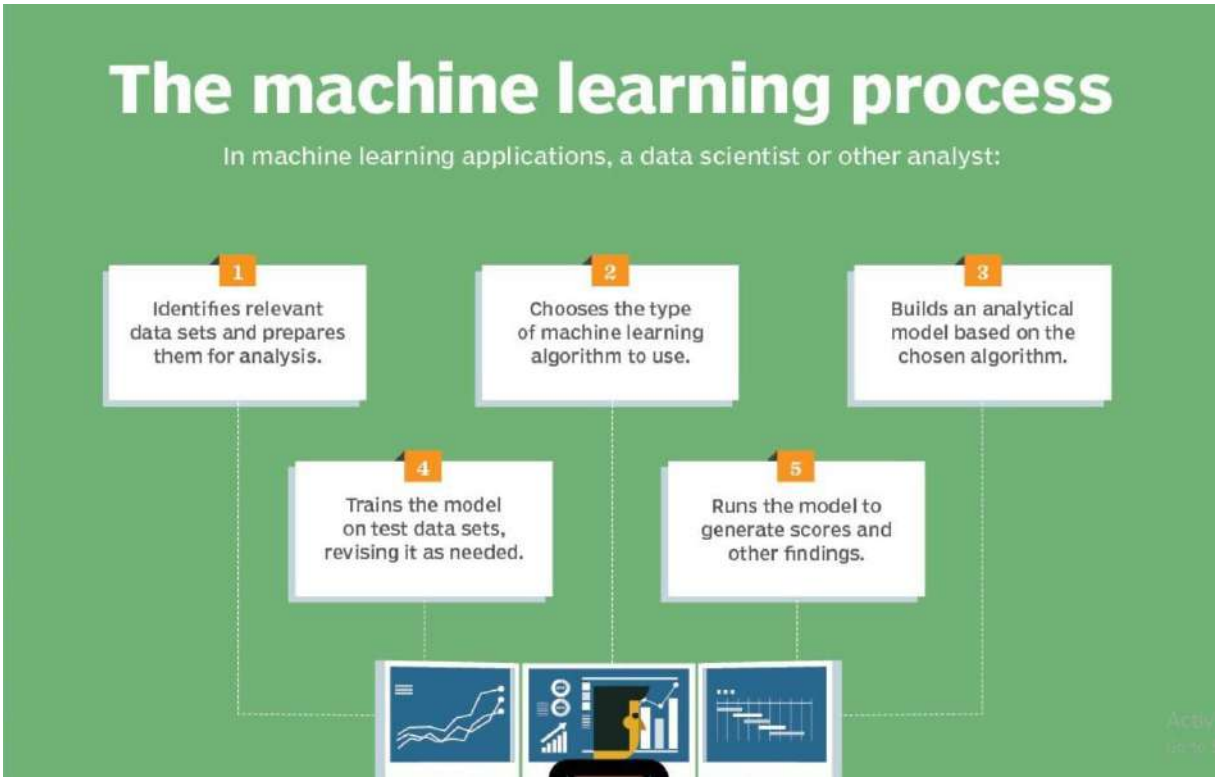
Processes of machine learning



There are three types of machine learning:

- Supervised Learning,
- Unsupervised Learning and
- Reinforcement Learning.

Machine Learning Algorithms

- Machine Learning algorithms are the programs that can learn the hidden patterns from the data, predict the output, and improve the performance from experiences on their own.

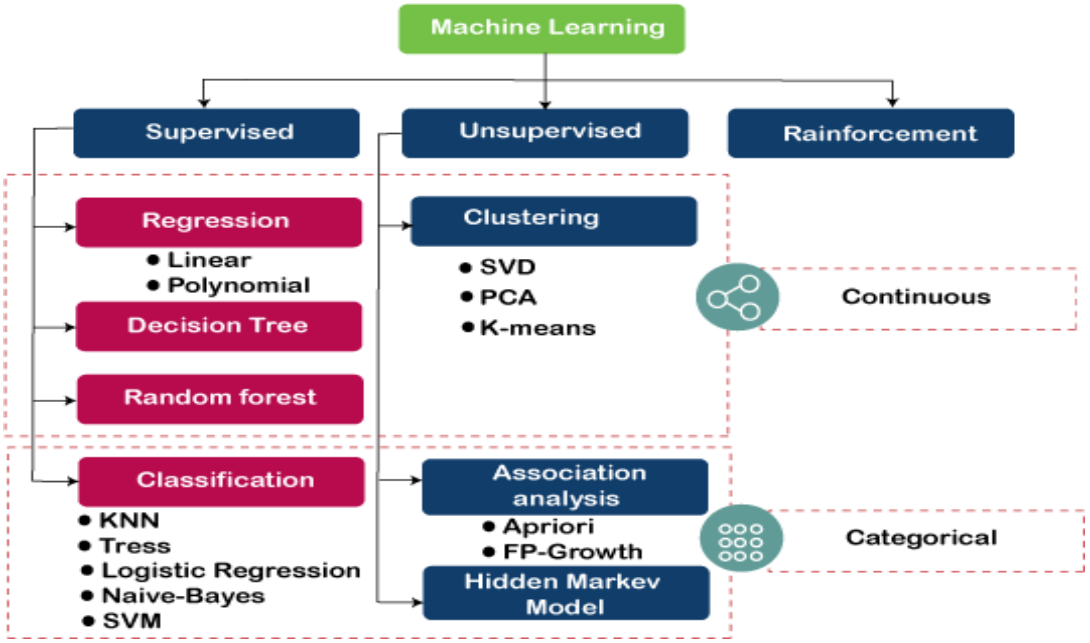
- Different algorithms can be used in machine learning for different tasks, such as simple linear regression that can be used for prediction problems like stock market prediction, and the KNN algorithm can be used for classification problems.
- In this topic, we will see the overview of some popular and most commonly used machine learning algorithms along with their use cases and categories.

Types of Machine Learning Algorithms:

Machine Learning Algorithm can be broadly classified into three types:

1. Supervised Learning Algorithms
2. Unsupervised Learning Algorithms
3. Reinforcement Learning algorithm

The below diagram illustrates the different ML algorithm, along with the categories:



1) Supervised Learning Algorithm:

- Supervised learning is a type of Machine learning in which the machine needs external supervision to learn.
- The supervised learning models are trained using the labelled dataset.

- Once the training and processing are done, the model is tested by providing a sample test data to check whether it predicts the correct output.
- The goal of supervised learning is to map input data with the output data.
- Supervised learning is based on supervision, and it is the same as when a student learns things in the teacher's supervision.
- The example of supervised learning is spam filtering.

Supervised learning can be divided further into two categories of problem:

- Classification
- Regression

2) Unsupervised Learning Algorithm :

- It is a type of machine learning in which the machine does not need any external supervision to learn from the data, hence called unsupervised learning.
- The unsupervised models can be trained using the unlabelled dataset that is not classified, nor categorized, and the algorithm needs to act on that data without any supervision.
- In unsupervised learning, the model doesn't have a predefined output, and it tries to find useful insights from the huge amount of data.
- These are used to solve the Association and Clustering problems.

Hence further, it can be classified into two types:

- Clustering
- Association

3) Reinforcement Learning:

In Reinforcement learning, an agent interacts with its environment by producing actions, and learn with the help of feedback.

List of Popular Machine Learning Algorithm

1. Linear Regression Algorithm
2. Logistic Regression Algorithm
3. Decision Tree
4. SVM
5. Naïve Bayes
6. KNN
7. K-Means Clustering
8. Random Forest
9. Apriori
10. PCA

1. Linear Regression:

- Linear regression is one of the most popular and simple machine learning algorithms that is used for predictive analysis.
- Here, predictive analysis defines prediction of something, and linear regression makes predictions for continuous numbers such as salary, age, etc.
- It shows the linear relationship between the dependent and independent variables, and shows how the dependent variable(y) changes according to the independent variable (x).
- It tries to best fit a line between the dependent and independent variables, and this best fit line is knowns as the regression line.

The equation for the regression line is:

$y= a_0+ a*x+ b$

Here, y= dependent variable

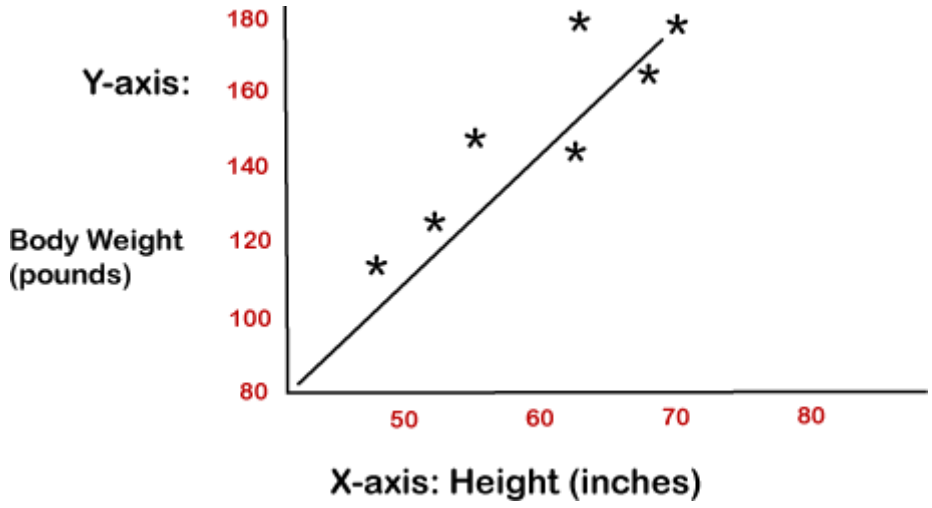
x= independent variable

a_0 = Intercept of line.

Linear regression is further divided into two types:

Simple Linear Regression: In simple linear regression, a single independent variable is used to predict the value of the dependent variable.

Multiple Linear Regression: In multiple linear regression, more than one independent variables are used to predict the value of the dependent variable.



2. Logistic Regression:

- Logistic regression is the supervised learning algorithm, which is used to predict the categorical variables or discrete values.
- It can be used for the classification problems in machine learning, and the output of the logistic regression algorithm can be either Yes or NO, 0 or 1, Red or Blue, etc.
- Logistic regression is similar to the linear regression except how they are used, such as Linear regression is used to solve the regression problem and predict continuous values, whereas Logistic regression is used to solve the Classification problem and used to predict the discrete values.
- Instead of fitting the best fit line, it forms an S-shaped curve that lies between 0 and 1.

3. Decision Tree Algorithm:

- A decision tree is a supervised learning algorithm that is mainly used to solve the classification problems but can also be used for solving the regression problems.

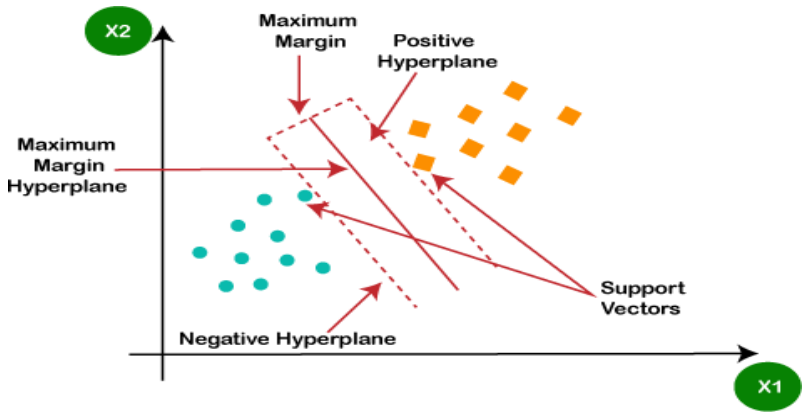
- It can work with both categorical variables and continuous variables.

- It shows a tree-like structure that includes nodes and branches, and starts with the root node that expand on further branches till the leaf node.
- The internal node is used to represent the features of the dataset, branches show the decision rules, and leaf nodes represent the outcome of the problem.

4. Support Vector Machine Algorithm:

- A support vector machine or SVM is a supervised learning algorithm that can also be used for classification and regression problems.
- However, it is primarily used for classification problems.
- The goal of SVM is to create a hyper plane or decision boundary that can segregate datasets into different classes.
- The data points that help to define the hyper plane are known as support vectors, and hence it is named as support vector machine algorithm.
- Some real-life applications of SVM are face detection, image classification, Drug discovery, etc.

Consider the below diagram:



5. Naïve Bayes Algorithm:

- Naïve Bayes classifier is a supervised learning algorithm, which is used to make predictions based on the probability of the object.
- The algorithm named as Naïve Bayes as it is based on Bayes theorem, and follows the naïve assumption that says' variables are independent of each other.

- The Bayes theorem is based on the conditional probability; it means the likelihood that event(A) will happen, when it is given that event(B) has already happened.

The equation for Bayes theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Naïve Bayes classifier is one of the best classifiers that provide a good result for a given problem.
- It is easy to build a naïve bayesian model, and well suited for the huge amount of dataset.

6. K-Nearest Neighbour (KNN):

- K-Nearest Neighbour is a supervised learning algorithm that can be used for both classification and regression problems.
- This algorithm works by assuming the similarities between the new data point and available data points.
- Based on these similarities, the new data points are put in the most similar categories.
- It is also known as the lazy learner algorithm as it stores all the available datasets and classifies each new case with the help of K-neighbours.
- The new case is assigned to the nearest class with most similarities, and any distance function measures the distance between the data points.

7. K-Means Clustering:

- K-means clustering is one of the simplest unsupervised learning algorithms, which is used to solve the clustering problems.
- The datasets are grouped into K different clusters based on similarities and dissimilarities, it means, datasets with most of the commonalties remain in one cluster which has very less or no commonalities between other clusters.
- In K-means, K-refers to the number of clusters, and means refer to the averaging the dataset in order to find the centroid.

- It is a centroid-based algorithm, and each cluster is associated with a centroid.
- This algorithm aims to reduce the distance between the data points and their centroids within a cluster.
- This algorithm starts with a group of randomly selected centroids that form the clusters at starting and then perform the iterative process to optimize these centroids' positions.

8. Random Forest Algorithm:

- Random forest is the supervised learning algorithm that can be used for both classification and regression problems in machine learning.
- It is an ensemble learning technique that provides the predictions by combining the multiple classifiers and improve the performance of the model.
- It contains multiple decision trees for subsets of the given dataset, and find the average to improve the predictive accuracy of the model.
- A random-forest should contain 64-128 trees. The greater number of trees leads to higher accuracy of the algorithm.
- To classify a new dataset or object, each tree gives the classification result and based on the majority votes, the algorithm predicts the final output.
- Random forest is a fast algorithm, and can efficiently deal with the missing & incorrect data.

9. Apriori Algorithm:

- Apriori algorithm is the unsupervised learning algorithm that is used to solve the association problems.
- It uses frequent item sets to generate association rules, and it is designed to work on the databases that contain transactions.
- With the help of these association rule, it determines how strongly or how weakly two objects are connected to each other.
- This algorithm uses a breadth-first search and Hash Tree to calculate the item set efficiently.
- The algorithm process iteratively for finding the frequent item sets from the large dataset.
- The apriori algorithm was given by the R. Agrawal and Srikant in the year 1994.

- It is mainly used for market basket analysis and helps to understand the products that can be bought together.
- It can also be used in the healthcare field to find drug reactions in patients.

10. Principle Component Analysis:

- Principle Component Analysis (PCA) is an unsupervised learning technique, which is used for dimensionality reduction.
- It helps in reducing the dimensionality of the dataset that contains many features correlated with each other.
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.
- It is one of the popular tools that is used for exploratory data analysis and predictive modelling.
- PCA works by considering the variance of each attribute because the high variance shows the good split between the classes, and hence it reduces the dimensionality.
- Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels.

Probability and Statistics Books for Machine Learning

- Probability and statistics both are the most important concepts for Machine Learning. Probability is about predicting the likelihood of future events, while statistics involves the analysis of the frequency of past events.
- Machine Learning has become one of the first choices for most freshers and IT professionals. But, in order to enter this field, one must have some pre-specified skills and one of those skills in Mathematics.
- Yes, Mathematics is very much important to learn ML technology and develop efficient applications for the business.
- When talking about mathematics for Machine Learning, it especially focuses on Probability and Statistics, which are the essential topics to get started with ML.
- Probability and statistics are considered as the base foundation for ML and data science to develop ML algorithms and build decision-making capabilities.
- Also, Probability and statistics are the primary prerequisites to learn ML.